A Robust Class of Data Languages and an Application to Learning

Benedikt Bollig, Peter Habermehl, Martin Leucker, Benjamin Monmege

Logical Methods in Computer Science
Vol. 10(4:19)2014, pp. 1–23
www.lmcs-online.org

Submitted Jan. 29, 2014
Published Dec. 29, 2014

A ROBUST CLASS OF DATA LANGUAGES AND AN APPLICATION TO LEARNING *

BENEDIKT BOLLIG [a], PETER HABERMEHL [b], MARTIN LEUCKER [c], AND BENJAMIN MONMEGE [d]

[a] LSV, ENS Cachan, CNRS & Inria, France
e-mail address: bollig@lsv.ens-cachan.fr

[b] Univ Paris Diderot, Sorbonne Paris Cité, LIAFA, CNRS, France
e-mail address: haberm@liafa.univ-paris-diderot.fr

[c] ISP, University of Lübeck, Germany
e-mail address: leucker@isp.uni-luebeck.de

[d] Université Libre de Bruxelles, Belgium
e-mail address: bmonmege@ulb.ac.be

Abstract. We introduce *session automata*, an automata model to process data words, i.e., words over an infinite alphabet. Session automata support the notion of *fresh* data values, which are well suited for modeling protocols in which sessions using fresh values are of major interest, like in security protocols or ad-hoc networks. Session automata have an expressiveness partly extending, partly reducing that of classical register automata. We show that, unlike register automata and their various extensions, session automata are robust: They (i) are closed under intersection, union, and (resource-sensitive) complementation, (ii) admit a symbolic regular representation, (iii) have a decidable inclusion problem (unlike register automata), and (iv) enjoy logical characterizations. Using these results, we establish a learning algorithm to infer session automata through membership and equivalence queries.

1. Introduction

The study of automata over data words, i.e., words over an infinite alphabet, has its origins in the seminal work by Kaminski and Francez [21]. Their finite-memory automata (more commonly called *register automata*) equip finite-state machines with registers in which data values (from the infinite alphabet) can be stored and be reused later. Register automata preserve some of the good properties of finite automata: they have a decidable emptiness problem and are closed under union and intersection. On the other hand, register automata are neither determinizable nor closed under complementation, and they have an undecidable equivalence/inclusion problem. There are actually several variants of register automata,

2012 ACM CCS: [**Theory of computation**]: Formal languages and automata theory—Automata over infinite objects; Theory and algorithms for application domains—Machine learning theory—Active learning.

Key words and phrases: Register Automata; Data words; Angluin-style learning; Freshness.

* This paper is an extended and revised version of the paper "A Fresh Approach to Learning Register Automata" which appeared in DLT 2013.

LOGICAL METHODS IN COMPUTER SCIENCE
DOI:10.2168/LMCS-10(4:19)2014

which all have the same expressive power but differ in the complexity of decision problems [14, 5]. In the sequel, many more automata models have been introduced (not necessarily with registers), aiming at a good balance between expressivity, decidability, and closure properties [29, 14, 23, 7, 17, 16]. Some of those models extend register automata, inheriting their drawbacks such as undecidability of the equivalence problem.

We will follow the work on register automata and study a model that supports the notion of *freshness*. When reading a data value, it may enforce that the data value is *fresh*, i.e., it has not occurred in the whole history of the run. This feature has been proposed in [33] to model computation with names in the context of programming-language semantics. Actually, fresh names are needed to model object creation in object-oriented languages, and they are important ingredients in modeling security protocols which often make use of so-called fresh nonces to achieve their security assertions [24]. Fresh names are also crucial in the field of network protocols, and they are one of the key features of the π-calculus [28]. Like ordinary register automata, fresh-register automata preserve some of the good properties of finite automata. However, they are not closed under complement and also come with an undecidable equivalence problem.

In this paper, we propose *session automata*, a robust automata model over data words. Like register automata, session automata are a syntactical restriction of fresh-register automata, but in an orthogonal way. Register automata drop the feature of checking *global freshness* (referring to the whole history) while keeping a local variant (referring to the registers). Session automata, on the other hand, discard local freshness, while keeping the global one. Session automata are well-suited whenever fresh values are important for a finite period, for which they will be stored in one of the registers. They correspond to the model from [8] without stacks.

Not surprisingly, we will show that session automata and register automata describe incomparable classes of languages of data words, whereas both are strictly weaker than fresh-register automata. Contrary to finite-state unification based automata introduced in [22], session automata (like fresh-register automata) do not have the capability to reset the content of a register. However, they can test global freshness which the model of [22] cannot. The *variable automata* from [16] do not employ registers, but rather use bound and free variables. However, variable automata are close to our model: they use a finite set of bound variables to track the occurrences of some data values, and a single free variable for all other data values (that must be different from data values tracked by bound variables). Contrary to our model, variable automata cannot test for global freshness, but we are not able to recognize the language of all data words, contrary to them.

In this paper, we show that session automata (i) are closed under intersection, union, and resource-sensitive complementation[1], (ii) have a unique canonical form (analogous to minimal deterministic finite automata), (iii) have a decidable equivalence/inclusion problem, and (iv) enjoy logical characterizations. Altogether, this provides a versatile framework for languages over infinite alphabets.

In a second part of the paper, we present an application of our automata model in the area of learning, where decidability of the equivalence problem is crucial. Learning automata deals with the inference of automata based on some partial information, for example samples, which are words that either belong to the accepted language or not. A popular framework is that of active learning defined by Angluin [2] in which a learner may consult a teacher for

[1] A notion similar to [25], but for a different model.

so-called membership and equivalence queries to eventually infer the automaton in question. Learning automata has many applications in computer science. Notable examples are the use in model checking [15] and testing [3]. See [26] for an overview.

While active learning of regular languages is meanwhile well understood and is supported by freely available libraries such as LearnLib [27] and libalf [10], extensions beyond plain regular languages are still an area of active research. Recently, automata dealing with potentially infinite data as basis objects have been studied. Seminal works in this area are that of [1, 20] and [19]. While the first two use abstraction and refinement techniques to cope with infinite data, the second approach learns a sub-class of register automata. Note that session automata are incomparable with the model from [19]. Thanks to their closure and decidability properties, a conservative extension of Angluin's classical algorithm will do for their automatic inference.

Outline. The paper is structured as follows. In Section 2 we introduce session automata. Section 3 presents the main tool allowing us to establish the results of this paper, namely the use of data words in symbolic normal form and the construction of a canonical session automaton. The section also presents some closure properties of session automata and the decidability of the equivalence problem. Section 4 gives logical characterizations of our model. In Section 5, we present an active learning algorithm for session automata. This paper is an extended version of [9].

2. Data Words and Session Automata

We let $\mathbb{N}$ be the set of natural numbers and $\mathbb{N}_{>0}$ be the set of non-zero natural numbers. In the following, we fix a non-empty finite alphabet Σ of *labels* and an infinite set D of *data values*. In examples, we usually use $D = \mathbb{N}$. A *data word* over Σ and D is a sequence $w = (a_1, d_1) \cdots (a_n, d_n)$ of pairs $(a_i, d_i) \in \Sigma \times D$. In other words, w is an element from $(\Sigma \times D)^*$. For $d \in \{d_1, \ldots, d_n\}$, we let $\mathit{first}_w(d)$ denote the position $j \in \{1, \ldots, n\}$ where d occurs for the first time, i.e., such that $d_j = d$ and there is no $k < j$ such that $d_k = d$. Accordingly, we define $\mathit{last}_w(d)$ to be the last position where d occurs.

An example data word over $\Sigma = \{a, b\}$ and $D = \mathbb{N}$ is given by $w = (a, 8)(b, 4)(a, 8)(c, 3)$ $(a, 4)(b, 4)(a, 9)$. We have $\mathit{first}_w(4) = 2$ and $\mathit{last}_w(4) = 6$.

This section recalls two existing automata models over data words – namely register automata, previously introduced in [21], and fresh-register automata, introduced in [33] as a generalization of register automata. Moreover, we introduce the new model of session automata, our main object of interest.

Register automata (initially called finite-memory automata) equip finite-state machines with registers in which data values can be stored and be read out later. Fresh-register automata additionally come with an oracle that can determine if a data value is *fresh*, i.e., has not occurred in the history of a run. Both register and fresh-register automata are closed under union and intersection, and they have a decidable emptiness problem. However, they are not closed under complementation, and their equivalence problem is undecidable, which limits their application in areas such as model checking and automata learning. Session automata, on the other hand, are closed under (resource-sensitive) complementation, and they have a decidable inclusion/equivalence problem.

Given a set $\mathcal{R}$, we let $\mathcal{R}^{\uparrow} \stackrel{\text{def}}{=} \{r^{\uparrow} \mid r \in \mathcal{R}\}$, $\mathcal{R}^{\odot} \stackrel{\text{def}}{=} \{r^{\odot} \mid r \in \mathcal{R}\}$, and $\mathcal{R}^{\circledast} \stackrel{\text{def}}{=} \{r^{\circledast} \mid r \in \mathcal{R}\}$. In the automata models that we are going to introduce, $\mathcal{R}$ will be the set of registers.

Transitions will be labeled with an element from $\mathcal{R}^{\circledast} \cup \mathcal{R}^{\odot} \cup \mathcal{R}^{\uparrow}$, which determines a register and the operation that is performed on it. More precisely, $r^{\circledast}$ writes a globally fresh value into r, $r^{\odot}$ writes a locally fresh value into r, and $r^{\uparrow}$ uses the value that is currently stored in r. For $\pi \in \mathcal{R}^{\circledast} \cup \mathcal{R}^{\odot} \cup \mathcal{R}^{\uparrow}$, we let $\mathsf{reg}(\pi) = r$ if $\pi \in \{r^{\circledast}, r^{\odot}, r^{\uparrow}\}$. Similarly,

$$\mathsf{op}(\pi) = \begin{cases} \circledast & \text{if } \pi \text{ is of the form } r^{\circledast} \\ \odot & \text{if } \pi \text{ is of the form } r^{\odot} \\ \uparrow & \text{if } \pi \text{ is of the form } r^{\uparrow}. \end{cases}$$

Definition 2.1 (Fresh-Register Automaton, cf. [33])**.** A *fresh-register automaton* (over Σ and D) is a tuple $\mathcal{A} = (S, \mathcal{R}, \iota, F, \Delta)$ where

- S is the non-empty finite set of *states*,
- $\mathcal{R}$ is the non-empty finite set of *registers*,
- $\iota \in S$ is the *initial state*,
- $F \subseteq S$ is the set of *final states*, and
- Δ is a finite set of *transitions*: each transition is a tuple of the form $(s, (a, \pi), s')$ where $s, s' \in S$ are the source and target state, respectively, $a \in \Sigma$, and $\pi \in \mathcal{R}^{\circledast} \cup \mathcal{R}^{\odot} \cup \mathcal{R}^{\uparrow}$. We call (a, π) the *transition label*.

For a transition $(s, (a, \pi), s') \in \Delta$, we also write $s \xrightarrow{(a,\pi)} s'$. When taking this transition, the automaton moves from state s to state s' and reads a symbol $(a, d) \in \Sigma \times D$. If $\pi = r^{\uparrow} \in \mathcal{R}^{\uparrow}$, then d is the data value that is currently stored in register r. If $\pi = r^{\circledast} \in \mathcal{R}^{\circledast}$, then d is some *globally fresh* data value, which has not been read in the *whole* history of the run; d is then written into register r. Finally, if $\pi = r^{\odot} \in \mathcal{R}^{\odot}$, then d is some *locally fresh* data value, which is *currently* not stored in the registers; it will henceforth be stored in register r.

Let us formally define the semantics of $\mathcal{A}$. A *configuration* is a triple $\gamma = (s, \tau, U)$ where $s \in S$ is the current state, $\tau : \mathcal{R} \rightharpoonup D$ is a partial mapping encoding the current register assignment, and $U \subseteq D$ is the set of data values that have been used so far. By $dom(\tau)$, we denote the set of registers r such that $\tau(r)$ is defined. Moreover, $\tau(\mathcal{R}) \stackrel{\text{def}}{=} \{\tau(r) \mid r \in dom(\tau)\}$. We say that γ is *final* if $s \in F$. As usual, we define a transition relation over configurations and let $(s, \tau, U) \xRightarrow{(a,d)} (s', \tau', U')$, where $(a, d) \in \Sigma \times D$, if there is a transition $s \xrightarrow{(a,\pi)} s'$ such that the following conditions hold:

(1) $\begin{cases} d = \tau(\mathsf{reg}(\pi)) & \text{if } \mathsf{op}(\pi) = \uparrow \\ d \notin \tau(\mathcal{R}) & \text{if } \mathsf{op}(\pi) = \odot \\ d \notin U & \text{if } \mathsf{op}(\pi) = \circledast, \end{cases}$

(2) $dom(\tau') = dom(\tau) \cup \{\mathsf{reg}(\pi)\}$ and $U' = U \cup \{d\}$,

(3) $\tau'(\mathsf{reg}(\pi)) = d$ and $\tau'(r) = \tau(r)$ for all $r \in dom(\tau) \setminus \{\mathsf{reg}(\pi)\}$.

A run of $\mathcal{A}$ on a data word $(a_1, d_1) \cdots (a_n, d_n) \in (\Sigma \times D)^*$ is a sequence

$$\gamma_0 \xRightarrow{(a_1,d_1)} \gamma_1 \xRightarrow{(a_2,d_2)} \cdots \xRightarrow{(a_n,d_n)} \gamma_n$$

for suitable configurations $\gamma_0, \ldots, \gamma_n$ with $\gamma_0 = (\iota, \emptyset, \emptyset)$ (here the partial mapping $\emptyset$ represents the mapping with empty domain). The run is *accepting* if γ_n is a final configuration. The *language* $L(\mathcal{A}) \subseteq (\Sigma \times D)^*$ of $\mathcal{A}$ is then defined as the set of data words for which there is an accepting run.

Note that fresh-register automata cannot distinguish between data words that are equivalent up to permutation of data values. More precisely, given $w, w' \in (\Sigma \times D)^*$, we

write $w \approx w'$ if $w = (a_1, d_1) \cdots (a_n, d_n)$ and $w' = (a_1, d'_1) \cdots (a_n, d'_n)$ such that, for all $i, j \in \{1, \ldots, n\}$, we have $d_i = d_j$ iff $d'_i = d'_j$. For instance, $(a, 4)(b, 2)(b, 4) \approx (a, 2)(b, 5)(b, 2)$. In the following, the equivalence class of a data word w wrt. $\approx$ is written $[w]_\approx$. We call $L \subseteq (\Sigma \times D)^*$ a *data language* if, for all $w, w' \in (\Sigma \times D)^*$ such that $w \approx w'$, we have $w \in L$ if, and only if, $w' \in L$. In particular, $L(\mathcal{A})$ is a data language for every fresh-register automaton $\mathcal{A}$.

We obtain natural subclasses of fresh-register automata when we restrict the transition labels $(a, \pi) \in \Sigma \times (\mathcal{R}^{\circledast} \cup \mathcal{R}^{\odot} \cup \mathcal{R}^{\uparrow})$ in the transitions.

Definition 2.2 (Register Automaton, [21])**.** A *register automaton* is a fresh-register automaton where every transition label is from $\Sigma \times (\mathcal{R}^{\odot} \cup \mathcal{R}^{\uparrow})$.

Like register automata, session automata are a syntactical restriction of fresh-register automata, but in an orthogonal way. Instead of local freshness, they include the feature of global freshness.

Definition 2.3 (Session Automaton)**.** A *session automaton* is a fresh-register automaton where every transition label is from $\Sigma \times (\mathcal{R}^{\circledast} \cup \mathcal{R}^{\uparrow})$.

We first compare the three models of automata introduced above in terms of expressive power.

Example 2.4. Consider the set of labels $\Sigma = \{\mathsf{req}, \mathsf{ack}\}$ and the set of data values $D = \mathbb{N}$, representing an infinite supply of process identifiers (pids). We model a simple (sequential) system where processes can approach a server and make a request, indicated by req, and where the server can acknowledge these requests, indicated by ack. More precisely, $(\mathsf{req}, p) \in \Sigma \times D$ means that the process with pid p performs a request, which is acknowledged when the system executes (ack, p).

Figure 1(a) depicts a register automaton that recognizes the language L_1 of data words verifying the following conditions:

- there are at most two open requests at a time;
- a process waits for an acknowledgment before making another request;
- every acknowledgment is preceded by a request;
- requests are acknowledged in the order they are received.

In the figure, an edge label of the form $(\mathsf{req}, r_i^{\odot} \vee r_i^{\uparrow})$ shall denote that there are two transitions, one labeled with $(\mathsf{req}, r_i^{\odot})$, and one labeled with $(\mathsf{req}, r_i^{\uparrow})$. Whereas a transition labeled with $(\mathsf{req}, r_i^{\odot})$ is taken when the current data value does not appear currently in the registers (but could have appeared before in the data word) and store it in r_i, transition labeled with $(\mathsf{req}, r_i^{\uparrow})$ simply checks that the current data is stored in register r_i. The automaton models a server that can store two requests at a time and will acknowledge them in the order they are received. For example, it accepts $(\mathsf{req}, 8)(\mathsf{req}, 4)(\mathsf{ack}, 8)(\mathsf{req}, 3)(\mathsf{ack}, 4)(\mathsf{req}, 8)(\mathsf{ack}, 3)(\mathsf{ack}, 8)$.

When we want to guarantee that, in addition, every process makes at most one request, we need the global freshness operator. Figure 1(b) hence depicts a session automaton recognizing the language L_2 of all the data words of L_1 in which every process makes at most one request. Notice that the transition from s_0 to s_1 is now labeled with $(\mathsf{req}, r_1^{\circledast})$, so that this transition can only be taken in case the current data value has never been seen before. We obtain $\mathcal{A}_2$ from $\mathcal{A}_1$ by replacing every occurrence of $r_i^{\odot} \vee r_i^{\uparrow}$ with $r_i^{\circledast}$. While $(\mathsf{req}, 8)(\mathsf{req}, 4)(\mathsf{ack}, 8)(\mathsf{req}, 3)(\mathsf{ack}, 4)(\mathsf{req}, 8)(\mathsf{ack}, 3)(\mathsf{ack}, 8)$ is no longer contained in L_2, $(\mathsf{req}, 8)$ $(\mathsf{req}, 4)(\mathsf{ack}, 8)(\mathsf{req}, 3)(\mathsf{ack}, 4)(\mathsf{ack}, 3)$ is still accepted.

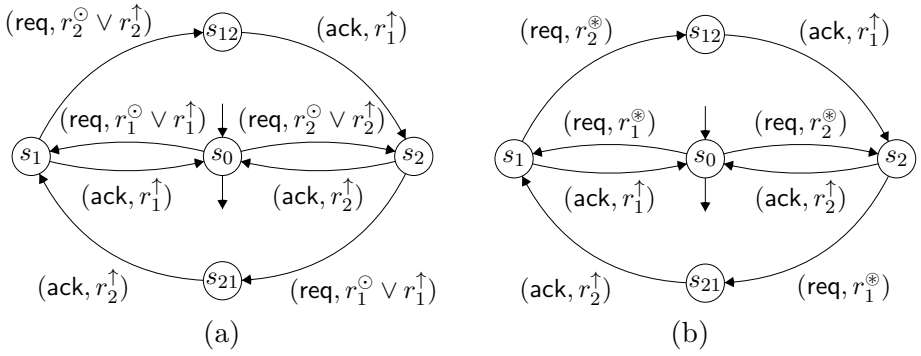


Figure 1: (a) Register automaton $\mathcal{A}_1$ for L_1, (b) Session automaton $\mathcal{A}_2$ for L_2

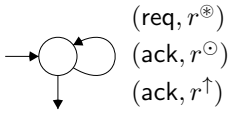


Figure 2: Fresh-register automaton $\mathcal{A}_3$ for L_3

As a last example, consider the language L_3 of data words in which every process makes at most one request (without any other condition). A fresh-register automaton recognizing it is given in Figure 2.

Proposition 2.5. *Register automata and session automata are incomparable in terms of expressive power. Moreover, fresh-register automata are strictly more expressive than both register automata and session automata.*

Proof. We use the languages L_1, L_2, and L_3 defined in Example 2.4 to separate the different automata models.

First, the language L_1, recognizable by a register automaton, is not recognized by any session automaton. Indeed, denoting w_d the data word $(\mathsf{req}, d)(\mathsf{ack}, d)$, no session automaton using k registers can accept

$$w_1 w_2 \cdots w_k w_{k+1} w_k \cdots w_2 w_1 \in L_1 \,.$$

Intuitively, the session automaton must store all $k+1$ data values of the requests in order to check the acknowledgement, and cannot discard any of the k first data values to store the $(k+1)$th since all of them have to be reused afterwards (and at that time they are not globally fresh anymore). More precisely, after reading $w_1 w_2 \cdots w_k$ the configuration must be of the form $(s, \tau, \{1, 2, \ldots, k\})$ with τ being a permutation of $\{1, \ldots, k\}$. Reading w_{k+1}, with fresh data value $k+1$, must then replace the content of a register with $k+1$. Suppose it is register j. Then, when reading the second occurrence of w_j, data value j is not globally fresh anymore, yet it is not stored anymore in the registers, which does not allow us to accept this data word.

Then, the language L_2, recognizable by a session automaton, is indeed not recognizable by a register automaton, for the same reasons as already developed in Proposition 5 of [21]. Intuitively, the automaton needs to register every data value encountered since it has to ensure the freshness of every pid.

Finally, language L_3, recognized by a fresh-register automaton, is not recognized by any register automaton (see again Proposition 5 of [21]) nor by any session automaton. In

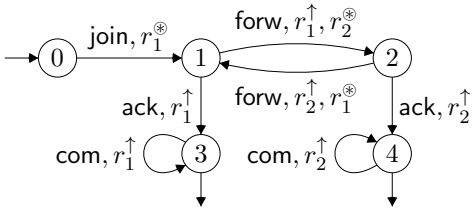


Figure 3: Session automaton for the P2P protocol

particular, no session automaton with k registers can accept the data word

$$(\mathsf{req}, 1)(\mathsf{req}, 2) \cdots (\mathsf{req}, k+1)(\mathsf{ack}, 1)(\mathsf{ack}, 2) \cdots (\mathsf{ack}, k+1) \in L_3$$

since when reading the letter $(\mathsf{req}, k+1)$, all the $k+1$ data values seen so far should be registered to accept the suffix afterwards. A formal proof can be done in the same spirit as for L_1. □

Example 2.6. To conclude the section, we present a session automaton with 2 registers that models a P2P protocol. A user can join a host with address x, denoted by action (join, x). The request is either forwarded by x to another host y, executing $(\mathsf{forw}_1, x)(\mathsf{forw}_2, y)$, or acknowledged by (ack, x). In the latter case, a connection between the user and x is established so that they can communicate, indicated by action (com, x). Note that the sequence of actions $(\mathsf{forw}_1, x)(\mathsf{forw}_2, y)$ should be considered as an encoding of a single action (forw, x, y) and is a way of dealing with actions that actually take two or more data values, as considered, e.g., in [19]. An example execution of our protocol is $(\mathsf{join}, 145)(\mathsf{forw}, 145, 978)(\mathsf{forw}, 978, 14)(\mathsf{ack}, 14)(\mathsf{com}, 14)(\mathsf{com}, 14)(\mathsf{com}, 14)$. In Figure 3, we show the session automaton for the P2P protocol: it uses 2 registers. Following [8], our automata can be easily extended to multi-dimensional data words. This also holds for the learning algorithm that will be presented in Section 5.

3. Symbolic Normal Form and Canonical Session Automata

Closure properties of session automata, decidability of inclusion/equivalence and the learning algorithm will be established by means of a symbolic normal form of a data word, as well as a canonical session automaton recognizing those normal forms. The crucial observation is that data equality in a data word recognized by a session automaton only depends on the transition labels that generate it. In this section, we suppose that the set of registers of a session automaton is of the form $\mathcal{R} = \{1, \ldots, k\}$. In the following, we let $\Gamma = \mathbb{N}_{>0}^{\circledast} \cup \mathbb{N}_{>0}^{\uparrow}$ and, for $k \geq 1$, $\Gamma_k = \{1, \ldots, k\}^{\circledast} \cup \{1, \ldots, k\}^{\uparrow}$.

3.1. Data Words in Symbolic Normal Forms.

Suppose a session automaton reads a sequence $u = (a_1, \pi_1) \cdots (a_n, \pi_n) \in (\Sigma \times \Gamma)^*$ of transition labels. We call u a *symbolic word*. It "produces" a data word if, and only if, a register is initialized before it is used. Formally, we say that u is *well-formed* if, for all positions $j \in \{1, \ldots, n\}$ with $\mathsf{op}(\pi_j) = \uparrow$, there is $i < j$ such that $\pi_i = \mathsf{reg}(\pi_j)^{\circledast}$. Let $\mathsf{WF} \subseteq (\Sigma \times \Gamma)^*$ be the set of all well-formed words.

With $u = (a_1, \pi_1) \cdots (a_n, \pi_n) \in (\Sigma \times \Gamma)^*$, we can associate an equivalence relation $\sim_u$ over $\{1, \ldots, n\}$, letting $i \sim_u j$ if, and only if,

- $\mathsf{reg}(\pi_i) = \mathsf{reg}(\pi_j)$, and

- $i \leq j$ and there is no position $k \in \{i+1, \ldots, j\}$ such that $\pi_k = \mathsf{reg}(\pi_i)^{\circledast}$, or
  $j \leq i$ and there is no position $k \in \{j+1, \ldots, i\}$ such that $\pi_k = \mathsf{reg}(\pi_j)^{\circledast}$.

If u is well-formed, then the data values of every data word $w = (a_1, d_1) \cdots (a_n, d_n)$ that a session automaton "accepts via" u conform with the equivalence relation $\sim_u$, that is, we have $d_i = d_j$ iff $i \sim_u j$. This motivates the following definition. Given a well-formed word $u = (a_1, \pi_1) \cdots (a_n, \pi_n) \in (\Sigma \times \Gamma)^*$, we call $w \in (\Sigma \times D)^*$ a *concretization* of u if it is of the form $w = (a_1, d_1) \cdots (a_n, d_n)$ such that, for all $i, j \in \{1, \ldots, n\}$, we have $d_i = d_j$ iff $i \sim_u j$. For example, $w = (a, 8)(a, 5)(b, 8)(a, 3)(b, 3)$ is a concretization of $u = (a, 1^{\circledast})(a, 2^{\circledast})(b, 1^{\uparrow})(a, 2^{\circledast})(b, 2^{\uparrow})$.

Let $\gamma(u)$ denote the set of all concretizations of u. Observe that, if w is a data word from $\gamma(u)$, then $\gamma(u) = [w]_{\approx}$. Concretization is extended to sets $L \subseteq (\Sigma \times \Gamma)^*$ of well-formed words, and we let $\gamma(L) \stackrel{\text{def}}{=} \bigcup_{u \in L \cap \mathsf{WF}} \gamma(u)$. Note that, here, we first filter the well-formed words before applying the operator. Now, let $\mathcal{A} = (S, \mathcal{R}, \iota, F, \Delta)$ be a session automaton. In the obvious way, we may consider $\mathcal{A}$ as a finite automaton over the finite alphabet $\Sigma \times (\mathcal{R}^{\circledast} \cup \mathcal{R}^{\uparrow})$. We then obtain a regular language $L_{symb}(\mathcal{A}) \subseteq (\Sigma \times \Gamma)^*$ (indeed, $L_{symb}(\mathcal{A}) \subseteq (\Sigma \times \Gamma_k)^*$ if $\mathcal{R} = \{1, \ldots, k\}$). It is not difficult to verify that $L(\mathcal{A}) = \gamma(L_{symb}(\mathcal{A}))$.

Though we have a symbolic representation of data languages recognized by session automata, it is in general difficult to compare their languages, since different symbolic words may give rise to the same concretizations. For example, we have $\gamma((a, 1^{\circledast})(a, 1^{\circledast})(a, 1^{\uparrow})) = \gamma((a, 1^{\circledast})(a, 2^{\circledast})(a, 2^{\uparrow}))$. However, we can associate, with every data word, a symbolic normal form, producing the same set of concretizations. Intuitively, the normal form uses the first (according to the natural total order) register whose current data value *will not be used anymore*. In the above example, $(a, 1^{\circledast})(a, 1^{\circledast})(a, 1^{\uparrow})$ would be in symbolic normal form: the data value stored at the first position in register 1 is not reused so that, at the second position, register 1 *must* be overwritten. For the same reason, $(a, 1^{\circledast})(a, 2^{\circledast})(a, 2^{\uparrow})$ is not in symbolic normal form, in contrast to $(a, 1^{\circledast})(a, 2^{\circledast})(a, 2^{\uparrow})(a, 1^{\uparrow})$ where register 1 is read at the end of the word.

Formally, given a data word $w = (a_1, d_1) \cdots (a_n, d_n)$, we define its symbolic normal form $\mathit{snf}(w) \stackrel{\text{def}}{=} (a_1, \pi_1) \cdots (a_n, \pi_n) \in (\Sigma \times \Gamma)^*$ inductively, along with sets $\mathit{Free}(i) \subseteq \mathbb{N}_{>0}$ indicating the registers that are reusable after executing position $i \in \{1, \ldots, n\}$. Setting $\mathit{Free}(0) = \mathbb{N}_{>0}$, we define

$$\pi_i = \begin{cases} \min(\mathit{Free}(i-1))^{\circledast} & \text{if } i = \mathit{first}_w(d_i) \\ \mathsf{reg}(\pi_{\mathit{first}_w(d_i)})^{\uparrow} & \text{otherwise}, \end{cases}$$

and

$$\mathit{Free}(i) = \begin{cases} \mathit{Free}(i-1) \setminus \min(\mathit{Free}(i-1)) & \text{if } i = \mathit{first}_w(d_i) \neq \mathit{last}_w(d_i) \\ \mathit{Free}(i-1) \cup \{\mathsf{reg}(\pi_i)\} & \text{if } i = \mathit{last}_w(d_i) \\ \mathit{Free}(i-1) & \text{otherwise}. \end{cases}$$

We canonically extend snf to data languages L, setting $\mathit{snf}(L) = \{\mathit{snf}(w) \mid w \in L\}$.

Example 3.1. Let $w = (a, 8)(b, 4)(a, 8)(c, 3)(a, 4)(b, 3)(a, 9)$. Then, we have $\mathit{snf}(w) = (a, 1^{\circledast})(b, 2^{\circledast})(a, 1^{\uparrow})(c, 1^{\circledast})(a, 2^{\uparrow})(b, 1^{\uparrow})(a, 1^{\circledast})$.

The relation between the mappings γ and snf is illustrated below

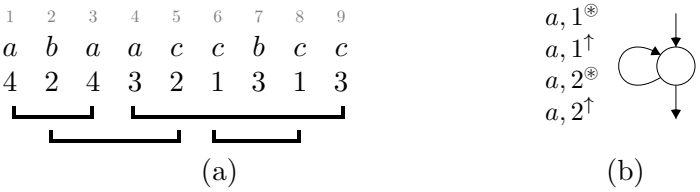
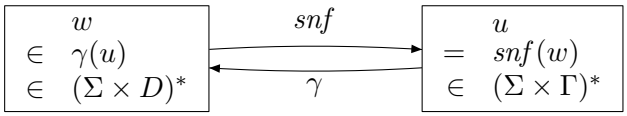


Figure 4: (a) A data word and its sessions, (b) Session automaton recognizing all 2-bounded data words



One easily verifies that $L = \gamma(\mathit{snf}(L))$, for all data languages L. Therefore, equality of data languages reduces to equality of their symbolic normal forms:

Lemma 3.2. *Let L and L' be data languages. Then, $L = L'$ if, and only if, $\mathit{snf}(L) = \mathit{snf}(L')$.*

Of course, symbolic normal forms may use any number of registers so that the set of symbolic normal forms is a language over an infinite alphabet as well. However, given a session automaton $\mathcal{A}$, the symbolic normal forms that represent the language $L(\mathcal{A})$ uses only a bounded (i.e., finite) number of registers. Indeed, an important notion in the context of session automata is the *bound* of a data word. Intuitively, the bound of $w = (a_1, d_1) \cdots (a_n, d_n) \in (\Sigma \times D)^*$ is the minimal number of registers that a session automaton needs in order to execute w. Or, in other words, the bound is the maximal number of overlapping *sessions*. A session is an interval delimiting the occurrence of one particular data value. Formally, a session of w is a set $I \subset \mathbb{N}_{>0}$ of the form $\{\mathit{first}_w(d), \mathit{first}_w(d) + 1, \ldots, \mathit{last}_w(d)\}$ with $d \in D$ a data value appearing in w. Given $k \in \mathbb{N}_{>0}$, we say that w is *k-bounded* if every position $i \in \{1, \ldots, n\}$ is contained in at most k sessions. Let DW_k denote the set of k-bounded data words, and let $\mathsf{SNF}_k \stackrel{\text{def}}{=} \mathit{snf}(\mathsf{DW}_k)$ denote the set of symbolic normal forms of all k-bounded data words.

One can verify that a data word w is k-bounded if, and only if, $\mathit{snf}(w)$ is a word over the alphabet $\Sigma \times \Gamma_k$. Notice that $\mathsf{DW}_k = \gamma((\Sigma \times \Gamma_k)^*)$. Indeed, inclusion $\mathsf{DW}_k \supseteq \gamma((\Sigma \times \Gamma_k)^*)$ is trivial. If, on the other hand, $w \in \mathsf{DW}_k$, we must have $\mathit{snf}(w) \in (\Sigma \times \Gamma_k)^*$, which implies that $w \in \gamma(\mathit{snf}(w)) \subseteq \gamma((\Sigma \times \Gamma_k)^*)$.

A data language L is said to be *k-bounded* if $L \subseteq \mathsf{DW}_k$. It is *bounded* if it is k-bounded for some k. Note that the set of all data words is not bounded.

Figure 4(a) illustrates a data word w with four different sessions. It is 2-bounded, as no position shares more than 2 sessions.

Example 3.3. Consider the session automaton from Figure 4(b). It recognizes the set of all 2-bounded data words over $\Sigma = \{a\}$.

3.2. **Deterministic Session Automata.** Session automata come with two natural notions of determinism. We call $\mathcal{A} = (S, \mathcal{R}, \iota, F, \Delta)$ *symbolically deterministic* if $|\{s' \in S \mid (s, (a, \pi), s') \in \Delta\}| \leq 1$ for all $s \in S$, $a \in \Sigma$, and $\pi \in \mathcal{R}^{\circledast} \cup \mathcal{R}^{\uparrow}$. Then, Δ can be seen as a partial function $S \times (\Sigma \times (\mathcal{R}^{\circledast} \cup \mathcal{R}^{\uparrow})) \rightharpoonup S$.

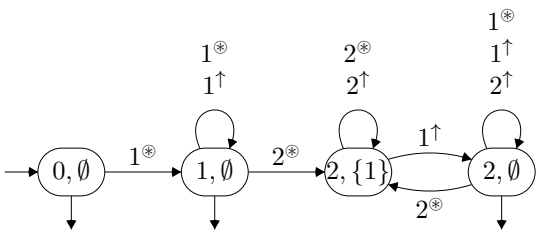


Figure 5: A session automaton recognizing SNF_2

We call $\mathcal{A}$ *data deterministic* if it is symbolically deterministic and, for all $s \in S$, $a \in \Sigma$, and $r_1, r_2 \in \mathcal{R}$ with $r_1 \neq r_2$, we have that $(s, (a, r_1^{\circledast})) \in dom(\Delta)$ implies $(s, (a, r_2^{\circledast})) \notin dom(\Delta)$. Intuitively, given a data word as input, the automaton is data deterministic if, in each state, given a pair letter/data value, there is at most one fireable transition.

Notice that session automata, even when symbolically or data deterministic, may not necessarily be "complete", in the sense that it is possible that a run over a data word falls into a deadlock situation: this is the case when the session automaton forced a data value to be removed from the set of registers, though it will be seen in the future.

While "data deterministic" implies "symbolically deterministic" by definition, the converse is not true. E.g., the session automaton $\mathcal{A}_2$ from Figure 1(b) and the one of Figure 4(b) are symbolically deterministic but not data deterministic. However, the session automaton obtained from $\mathcal{A}_2$ by removing, e.g., the transition from s_0 to s_2 (coupled with the transition from s_0 to s_1, it causes non-determinism when reading a fresh data value at a request), is data deterministic (and is indeed equivalent to $\mathcal{A}_2$, in the sense that it recognizes the same language $L(\mathcal{A}_2)$).

Example 3.4. We explain how to construct a symbolically deterministic session automaton $\mathcal{A}$, with $k \geq 1$ registers, such that $L_{symb}(\mathcal{A}) = \mathsf{SNF}_k$. Its state space is $S = \{0, \ldots, k\} \times 2^{\{1,\ldots,k\}}$, consisting of (i) the greatest register already initialized (indeed we will only use a register r if every register $r' < r$ has already been used), (ii) a subset P of registers that we promise to reuse again before resetting their value. The initial state of $\mathcal{A}$ is $(0, \emptyset)$, whereas the set of accepting states is $(\{0, \ldots, k\}) \times \{\emptyset\}$. We now describe the set of transitions. For every $a \in \Sigma$, $i \in \{0, \ldots, k\}$, $P \subseteq \{1, \ldots, k\}$, and $r \in \{1, \ldots, k\}$:

$$\Delta\big((i, P), (a, r^{\uparrow})\big) = \begin{cases} (i, P \setminus \{r\}) & \text{if } r \leq i \\ \text{not defined} & \text{otherwise} \end{cases}$$

$$\Delta\big((i, P), (a, r^{\circledast})\big) = \begin{cases} (\max(i, r), P \cup \{1, \ldots, r-1\}) & \text{if } r - 1 \leq i \wedge r \notin P \\ \text{not defined} & \text{otherwise} \end{cases}$$

Figure 5 depicts the session automaton for SNF_2 (omitting Σ).

By determinizing a finite-state automaton recognizing the symbolic language, it is easy to show that every language recognized by a session automaton is also recognized by a symbolically deterministic session automaton: we shall study this question in more detail in the next section. The next theorem shows that this is not true for data deterministic session automata.

Theorem 3.5. *Session automata are strictly more expressive than data deterministic session automata.*

Proof. We show that the data language $L = \mathsf{DW}_2$ cannot be recognized by a data deterministic session automaton. Indeed, suppose that such an automaton exists, with k registers. Then, consider the word $w = (a,1)(a,2)(a,3)\cdots(a,k+1) \in L$, where every data value is fresh. By data determinism, there is a unique run accepting w. Along this run, let $i < j$ be two positions such that their two fresh data values have been stored in the same register r (such a pair must exist since the automaton has only k registers). Without loss of generality, we can consider the greatest position j verifying this condition, and then the greatest position i associated with j. This means that register r is used for the last time when reading j, and has not been used in-between positions i and j. Now, the word $(a,1)(a,2)(a,3)\cdots(a,k+1)(a,i) \in L$ must be recognized by the automaton, but cannot since data value i appearing on the last position is not fresh anymore, and yet not stored in one of the registers (since register r was reused at j). □

3.3. **Canonical Session Automata.** We now present the main result of this section showing that every session automaton $\mathcal{A}$ is equivalent to a *canonical* session automaton $\mathcal{A}^C$, whose symbolic language $L_{symb}(\mathcal{A}^C)$ contains only symbolic normal forms.

Theorem 3.6. *Let $\mathcal{A} = (S, \mathcal{R}, \iota, F, \Delta)$ be a session automaton with $\mathcal{R} = \{1,\dots,k\}$. Then, $L(\mathcal{A})$ is k-bounded. Moreover, $snf(L(\mathcal{A}))$ is a regular language over the finite alphabet $\Sigma \times \Gamma_k$. A corresponding automaton $\tilde{\mathcal{A}}$ can be effectively computed. Its number of states is at most exponential in k and linear in $|S|$.*

Proof. First, if $\mathcal{A}$ is a session automaton using k registers, the language $L(\mathcal{A})$ is k-bounded since $L_{symb}(\mathcal{A}) \subseteq (\Sigma \times \Gamma_k)^*$, which implies that $L(\mathcal{A}) = \gamma(L_{symb}(\mathcal{A})) \subseteq \gamma((\Sigma \times \Gamma_k)^*) = \mathsf{DW}_k$.

Example 3.4, constructing a symbolically deterministic session automaton for $\mathsf{SNF}_k = snf(\gamma((\Sigma \times \Gamma_k)^*))$, shows that regularity of the symbolic language $(\Sigma \times \Gamma_k)^*$ is preserved under the application of $snf(\gamma(\cdot))$. We now prove that this is the case for every regular language over $\Sigma \times \Gamma_k$. In particular, for the symbolic *regular* language $L_{symb}(\mathcal{A})$, this will show that $snf(L(\mathcal{A}))$, which is equal to $snf(\gamma(L_{symb}(\mathcal{A})))$, is regular.

Let $L \subseteq (\Sigma \times \Gamma_k)^*$ be regular. Consider first the language

$$\tilde{L} = \{u \in \mathsf{WF} \cap (\Sigma \times \Gamma_k)^* \mid \text{there is } u' \in L \text{ such that } \gamma(u) = \gamma(u')\}$$

i.e., the set of well-formed symbolic words having the same concretizations as some word from L. We show that $snf(\gamma(L)) = \mathsf{SNF}_k \cap \tilde{L}$. Indeed, if $u \in snf(\gamma(L))$, then there are $u' \in L$ and $w \in \gamma(u')$ such that $u = snf(w)$. Since $u' \in (\Sigma \times \Gamma_k)^*$, we have $u \in snf(\gamma((\Sigma \times \Gamma_k)^*)) = \mathsf{SNF}_k$. Moreover, we have $[w]_{\approx} = \gamma(u')$ and $w \in \gamma(snf(w)) = \gamma(u)$ implying also $[w]_{\approx} = \gamma(u)$. Finally, we obtain $\gamma(u) = \gamma(u')$, so that $u \in \tilde{L}$. Reciprocally, if $u \in \mathsf{SNF}_k \cap \tilde{L}$, then there is $u' \in L$ such that $\gamma(u) = \gamma(u')$. Hence, starting from a word w in $\gamma(u)$ (which is non empty since u is well-formed), we have $u = snf(w)$ (by uniqueness of the symbolic normal form) and $w \in \gamma(u') \subseteq \gamma(L)$, so that $u \in snf(\gamma(L))$.

We know from Example 3.4 that SNF_k is regular. We now show that $\tilde{L}$ is regular: knowing that $snf(\gamma(L)) = \mathsf{SNF}_k \cap \tilde{L}$, this will permit to conclude that $snf(\gamma(L))$ is regular. To do so, let $\mathcal{A} = (S, \mathcal{R}, \iota, F, \Delta)$ be a session automaton with $\mathcal{R} = \{1,\dots,k\}$ such that $L_{symb}(\mathcal{A}) = L$. We construct a session automaton $\tilde{\mathcal{A}} = (S \times \mathrm{Inj}(k), \mathcal{R}, (s_0, \emptyset), F \times \mathrm{Inj}(k), \tilde{\Delta})$ recognizing the symbolic language $\tilde{L}$. Hereby, $\mathrm{Inj}(k)$ is the set of partial injective mappings from $\{1,\dots,k\}$ to $\{1,\dots,k\}$, and $\emptyset \in \mathrm{Inj}(k)$ denotes the mapping with empty domain. These partial mappings are used to remember the correspondence between old registers and

new ones, so they may be understood as a set of constraints. For example, the mapping $(2 \mapsto 1, 1 \mapsto 3)$ stands for "old register 2 henceforth refers to 1, and old register 1 henceforth refers to 3". Each subset of these constraints forms always a *valid* partial injective mapping. In the following, such a subset is called a sub-mapping. For example, $\sigma = (1 \mapsto 3)$ is a sub-mapping of the previous one; it can then be extended with the new constraint $2 \mapsto 2$, which we denote $\sigma[2 \mapsto 2]$. We describe now the transition relation of $\tilde{\mathcal{A}}$:

$$\begin{aligned}\tilde{\Delta} = \quad & \big\{\big((s_1,\sigma),(a,\sigma(r)^{\uparrow}),(s_2,\sigma)\big) \mid \big(s_1,(a,r^{\uparrow}),s_2\big) \in \Delta\big\} \\ \cup \, & \big\{\big((s_1,\sigma_1),(a,r_2^{\circledast}),(s_2,\sigma_2)\big) \mid \big(s_1,(a,r_1^{\circledast}),s_2\big) \in \Delta \wedge \sigma_2 = \sigma[r_1 \mapsto r_2] \\ & \qquad \text{with } \sigma \text{ maximal sub-mapping of } \sigma_1 \text{ s.t. } \sigma[r_1 \mapsto r_2] \text{ injective}\big\}\end{aligned}$$

We simulate $r^{\uparrow}$-transitions simply using the current mapping σ. For $r^{\circledast}$-transitions, we update σ, recording the new permutation of the registers: the maximal sub-mapping σ of σ_1 is either σ_1 itself or σ_1 where exactly one constraint $r_1 \mapsto r_3$ is removed to free r_1. One can indeed show that $L_{symb}(\tilde{\mathcal{A}}) = \tilde{L}$. Inclusion $L_{symb}(\tilde{\mathcal{A}}) \subseteq \tilde{L}$ is easy to show since an accepting run in $\tilde{\mathcal{A}}$ can be mapped to an accepting run in $\mathcal{A}$ using the partial injective mappings maintained in the states of $\tilde{\mathcal{A}}$. For the other inclusion, it suffices to prove that for every symbolic word $u \in L$ and well-formed word u' such that $\gamma(u') = \gamma(u)$, we have $u' \in L_{symb}(\tilde{\mathcal{A}})$. By definition of γ, we know that projections of u and u' over the finite alphabet Σ are the same, and that $\sim_u = \sim_{u'}$: the latter permits to reconstruct by induction a unique sequence of partial injective mappings linking the registers used in u and in u'. An accepting run of $\mathcal{A}$ on u can therefore be mapped to an accepting run of $\tilde{\mathcal{A}}$ on u'.

Building the product of the automaton recognizing SNF_k and the automaton $\tilde{\mathcal{A}}$, we obtain a session automaton using k registers recognizing $snf(\gamma(L))$. Its number of states is bounded above by $O(|Q| \times k! \times (k+1) \times 2^k)$ (as the number of partial injective mappings in $\mathrm{Inj}(k)$ is bounded above by $O(k!)$). □

From the automaton $\tilde{\mathcal{A}}$ built in the proof of the previous theorem, we can consider the (unique up to isomorphism) minimal deterministic finite-state automaton $\mathcal{A}^C$ (i.e., symbolically deterministic session automaton) equivalent to it: this automaton will be called the *canonical session automaton*. In case $\mathcal{A}$ is data deterministic, we can verify that $\tilde{\mathcal{A}}$ is symbolically deterministic, and hence the minimal automaton $\mathcal{A}^C$ has at most $O(|Q| \times k! \times (k+1) \times 2^k)$ states. Otherwise, a determinization phase has to be done resulting in a canonical session automaton with at most $2^{O(|Q| \times k! \times (k+1) \times 2^k)}$ states.

Example 3.7. Examples of $\mathcal{A}$ and $\tilde{\mathcal{A}}$, as defined in the previous proof, are given in Figure 6. The figure also depicts the canonical automaton $\mathcal{A}^C$ associated with $\mathcal{A}$, obtained by determinizing and minimizing the product of both $\tilde{\mathcal{A}}$ and the symbolically deterministic automaton recognizing SNF_2 (as given in Figure 5). Note that $\mathcal{A}^C$ is symbolically deterministic and minimal.

3.4. **Closure Properties.** Using Theorem 3.6, we obtain some language theoretical closure properties of session automata, which they inherit from classical regular languages. These results demonstrate a certain robustness as required in verification tasks such as compositional verification [11] and infinite-state regular model checking [18].

Theorem 3.8. *We have the following closure properties:*

- *Session automata are closed under union and intersection.*

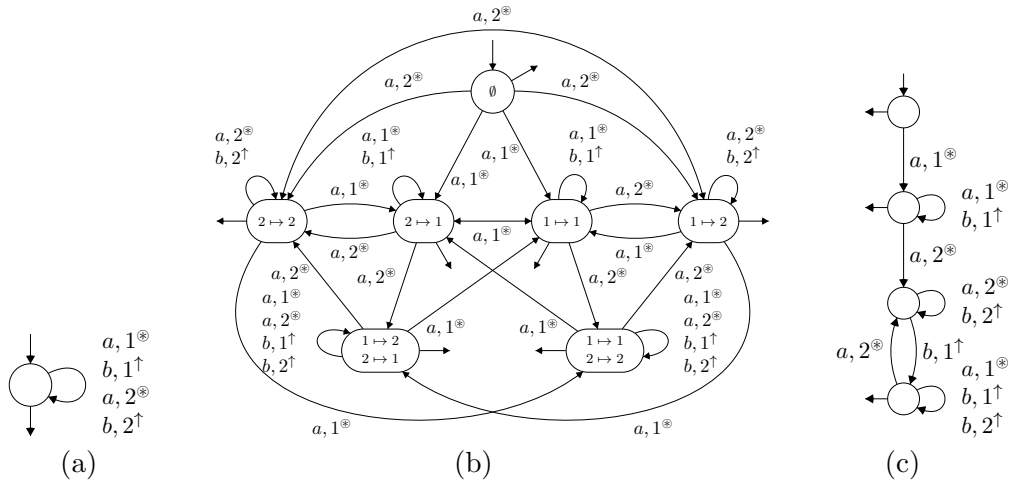


Figure 6: (a) A session automaton $\mathcal{A}$, (b) its automaton $\tilde{\mathcal{A}}$, (c) its canonical automaton $\mathcal{A}^C$

- *Session automata are closed under resource-sensitive complementation: Given a session automaton $\mathcal{A}$ with k registers, there is a session automaton $\mathcal{A}'$ with k registers such that $L(\mathcal{A}') = \mathsf{DW}_k \setminus L(\mathcal{A})$.*

Proof. Let $\mathcal{A}$ be a session automaton using k registers, and $\mathcal{B}$ a session automaton using k' registers. Using a classical product construction for $\mathcal{A}^C$ and $\mathcal{B}^C$, we obtain a session automaton using $\min(k, k')$ registers recognizing the data language $L(\mathcal{A}) \cap L(\mathcal{B})$. The language $L(\mathcal{A}) \cup L(\mathcal{B})$ is recognized by the session automaton, using $\max(k, k')$ registers, that we obtain as the "disjoint union" of $\mathcal{A}$ and $\mathcal{B}$, branching on the first transition in one of these two automata.

Finally, let us consider a symbolically deterministic session automaton $\mathcal{A}$ using k registers. Without loss of generality, by adding a sink state, we can suppose that $\mathcal{A}$ is complete. Then, every well-formed symbolic word over $\Sigma \times \Gamma_k$ has exactly one run in $\mathcal{A}$. The automaton $\mathcal{A}'$ constructed from $\mathcal{A}$ by taking as accepting states the non-accepting states of $\mathcal{A}$ verifies that $L_{symb}(\mathcal{A}') = (\Sigma \times \Gamma_k)^* \setminus L_{symb}(\mathcal{A})$ so that $L(\mathcal{A}') = \gamma((\Sigma \times \Gamma_k)^*) \setminus L(\mathcal{A})$. Notice that $\mathcal{A}'$ is symbolically deterministic, but not necessarily data deterministic (even if $\mathcal{A}$ is), because of the completion step. □

Theorem 3.9. *The inclusion problem for session automata is decidable.*

Proof. Considering two session automata $\mathcal{A}$ and $\mathcal{B}$, we can decide inclusion $L(\mathcal{A}) \subseteq L(\mathcal{B})$ by considering the canonical automata $\mathcal{A}^C$ and $\mathcal{B}^C$. Indeed, $L(\mathcal{A}) \subseteq L(\mathcal{B}) \iff \mathit{snf}(L(\mathcal{A})) \subseteq \mathit{snf}(L(\mathcal{B})) \iff L_{symb}(\mathcal{A}^C) \subseteq L_{symb}(\mathcal{B}^C)$. Thus, it is sufficient to check inclusion for $\mathcal{A}^C$ and $\mathcal{B}^C$. □

In case $\mathcal{B}$ is data deterministic, $\mathcal{B}^C$ has a size polynomial in the number of states of $\mathcal{B}$, but exponential in the number of registers. Testing the inclusion $L_{symb}(\mathcal{A}^C) \subseteq L_{symb}(\mathcal{B}^C)$ may be done by first complementing $L_{symb}(\mathcal{B}^C)$ (which does not add states since $\mathcal{B}^C$ is symbolically deterministic) and then testing the emptiness of its intersection with $L_{symb}(\mathcal{A}^C)$. In the overall, this implies a complexity of the inclusion check that is polynomial in the number of states of $\mathcal{A}$ and $\mathcal{B}$, but exponential in the number of registers used by $\mathcal{B}$. In case

$\mathcal{B}$ is not data deterministic, a determinization phase may add an exponent in the size and the number of registers of $\mathcal{B}$.

As a corollary, we obtain that the emptiness problem and the universality problem under k-boundedness (i.e., knowing whether the language of a session automaton with k registers is the whole set of k-bounded data words) are decidable for session automata. It is not surprising for the emptiness problem, since it already holds for fresh-register automata. Notice that the problem is shown co-NP-complete for register automata in [31], and we can show that the emptiness problem is co-NP-complete, too, for session automata. First, co-NP-hardness can be shown by a reduction to the 3-SAT problem, in a very similar way as in [31]. Then, the co-NP upper bound comes from the symbolic view. Indeed, for a session automaton $\mathcal{A}$ with k registers, $L(\mathcal{A}) = \emptyset$ if and only if $L_{symb}(\mathcal{A}) \cap \mathsf{WF}_k = \emptyset$ (where WF_k denotes the set of well-formed symbolic words over alphabet $\Sigma \times \Gamma_k$). We may not construct a finite automaton recognizing $L_{symb}(\mathcal{A}) \cap \mathsf{WF}_k$ (that has a size exponential in k), but instead non-deterministically search for a witness of non-emptiness of $L_{symb}(\mathcal{A}) \cap \mathsf{WF}_k$, i.e., a well-formed word u such that $u \in L_{symb}(\mathcal{A})$. Notice that the membership test of u in the finite-state automaton $\mathcal{A}$ can be performed in polynomial time, hence, to conclude, we must simply show the existence of a well-formed witness u of polynomial size. As for register automata in [31], this relies on the fact that, even though the total number of configurations of $\mathcal{A}$ is exponential in k (due to the set U of initialized registers), along a run of $\mathcal{A}$, only a polynomial (in $\mathcal{A}$ and k) number of configurations can be visited, since the set U will take at most $k+1$ values during the computation (the initialization of registers is done in a certain order, and no register can be emptied at any point). Hence, by disallowing the visit of two occurrences of the same configuration, the existence of a well-formed witness implies the existence of a well-formed witness of polynomial size.

While fresh-register automata are not complementable for the set of all data words, they are complementable for k-bounded data words, using the previous theorem. The reason is that, given a fresh-register automaton $\mathcal{A}$, one can construct a session automaton $\mathcal{B}$ such that $L(\mathcal{B}) = L(\mathcal{A}) \cap \mathsf{DW}_k$.

4. Logical Characterizations

In this section, we provide logical characterizations of session automata.

4.1. MSO Logic over Data Words.

We consider the standard *data monadic second-order logic* (dMSO), which is an extension of classical MSO logic by the binary predicate $x \sim y$ to compare data values.

We fix infinite supplies of first-order variables $x, y, \ldots$, which are interpreted as word positions, and second-order variables $X, Y, \ldots$, which are taken as sets of positions. We let dMSO be the set of formulae φ defined by the following grammar:

$$\varphi \ ::= \ \mathrm{label}(x) = a \mid x = y \mid y = x + 1 \mid x \sim y \mid x \in X \mid \neg\varphi \mid \varphi \vee \varphi \mid \exists x\, \varphi \mid \exists X\, \varphi$$

with x, y first-order variables, X a second-order variable and $a \in \Sigma$. The semantics of formulae in dMSO is given in Table 1: we define $w, \sigma \models \varphi$ (to be read as "w satisfies φ when free variables of φ are interpreted as prescribed in σ") by induction over φ, where $w = (a_1, d_1) \cdots (a_n, d_n) \in (\Sigma \times D)^*$ is a data word and σ is a valuation of (at least the) free variables in φ, i.e., such that for every first-order free variable x, we have $\sigma(x) \in \{1, \ldots, n\}$ and for every second-order free variable X, we have $\sigma(X) \subseteq \{1, \ldots, n\}$. For a first-order

Table 1: Semantics of formulae in dMSO

$w, [x \mapsto i, y \mapsto j] \models x = y$	if $i = j$
$w, [x \mapsto i] \models \text{label}(x) = a$	if $a_i = a$
$w, [x \mapsto i, y \mapsto j] \models y = x + 1$	if $j = i + 1$
$w, [x \mapsto i, X \mapsto I] \models x \in X$	if $i \in I$
$w, [x \mapsto i, y \mapsto j] \models x \sim y$	if $d_i = d_j$
$w, \sigma \models \neg\varphi$	if $w, \sigma \not\models \varphi$
$w, \sigma \models \varphi_1 \vee \varphi_2$	if $w, \sigma \models \varphi_1$ or $w, \sigma \models \varphi_2$
$w, \sigma \models \exists x\ \varphi$	if there exists $i \in \{1, \dots, n\}$ such that $w, \sigma[x \mapsto i] \models \varphi$
$w, \sigma \models \exists X\ \varphi$	if there exists $I \subseteq \{1, \dots, n\}$ such that $w, \sigma[X \mapsto I] \models \varphi$

variable x and a position $i \in \{1, \dots, n\}$, we let $\sigma[x \mapsto i]$ be the valuation τ such that $\tau(x) = i$ and $\tau(\alpha) = \sigma(\alpha)$ for every variable α different from x. A similar definition holds for second-order variables.

In addition, we use abbreviations such as *true*, $x \leq y$, $\forall x\ \varphi$, $\varphi \wedge \psi$, $\varphi \to \psi$, etc. A sentence is a formula without free variables. For a dMSO sentence φ, we set $L(\varphi) \stackrel{\text{def}}{=} \{w \in (\Sigma \times D)^* \mid w \models \varphi\}$. Note that $L(\varphi)$ is a data language.

As usual, to deal with free variables, it is possible to extend the alphabet Σ as follows. If V is the set of variables that occur in φ, we have to consider data words over $\hat{\Sigma} = \Sigma \times \{0, 1\}^V$ and D. Intuitively, these data words include the interpretation of the free variables. If a data word carries, at position i, the letter $(a, \bar{b}, d) \in \hat{\Sigma} \times D$ with $\bar{b}[x] = 1$ (where $\bar{b}[x]$ refers to the x-component of $\bar{b}$), then x is interpreted as position i. If $\bar{b}[X] = 1$, then X is interpreted as a set *containing* i. Whenever we refer to a word over the extended alphabet $\hat{\Sigma}$, we will silently assume that the interpretation of a first-order variable x is uniquely determined, i.e., there is exactly one position i where $\bar{b}[x] = 1$. This is justified, since the set of those "well-shaped" words is (symbolically) regular. This way we can transform any well-shaped word $\hat{w} \in (\Sigma \times \{0, 1\}^V \times D)^*$ into a pair (w, σ) where w is a data word of $(\Sigma \times D)^*$ and σ is a valuation of variables in V, and vice versa.

Note that dMSO is a very expressive logic and goes beyond virtually all automata models defined for data words [29, 32, 6, 12]. However, if we restrict to bounded languages, we can show that dMSO is no more expressive than session automata.

Theorem 4.1. *Let L be a bounded data language. Then, the following statements are equivalent:*

- *There is a session automaton $\mathcal{A}$ such that $L(\mathcal{A}) = L$.*
- *There is a dMSO sentence φ such that $L(\varphi) = L$.*

Proof. The construction of a dMSO formula of the form $\exists X_1 \cdots \exists X_m\ (\alpha \wedge \forall x \forall y\ (x \sim y \leftrightarrow \beta))$, with α and β classical MSO formulae (not containing predicate $\sim$), from a session automaton $\mathcal{A}$ was implicitly shown in [8] (with a different goal, though). The idea is that the existential second-order variables $X_1, \dots, X_m$ are used to guess an assignment of transitions to positions. In α, it is verified that the assignment corresponds to a run of $\mathcal{A}$. Moreover, β checks if data equality corresponds to the data flow as enforced by the transition labels from Γ_k. The formula has a size polynomial in the size of the automaton. In Section 4.2, formulae of this shape will be studied in more detail.

For the converse direction, we perform, as usual, an induction on the structure of the formula φ of dMSO such that $L(\varphi)$ is k-bounded, for some $k \geq 1$. To cope with free variables, we use the encoding of a pair (w, σ) as presented before.

First, we have to deal with the base cases:

- Consider the formula $\mathrm{label}(x) = a$. We construct a session automaton $\mathcal{A}$ with k registers such that $L_{symb}(\mathcal{A})$ consists of all "well-shaped" words $u \in (\hat{\Sigma} \times \Gamma_k)^*$ containing a letter $(a, \bar{b}, \pi)$ with $\bar{b}[x] = 1$.
- For $x = y$, the automaton has to verify that there is a letter $(a, \bar{b}, \pi)$ such that $\bar{b}[x] = \bar{b}[y] = 1$, which can be done since this is a regular condition on word over alphabet $\hat{\Sigma} \times \Gamma_k$.
- Formulae $y = x + 1$ and $x \in X$ are treated similarly.
- The most interesting base case is $x \sim y$. Let L be the symbolic language containing exactly the symbolic words $u = (a_1, \bar{b}_1, \pi_1) \cdots (a_n, \bar{b}_n, \pi_n) \in (\hat{\Sigma} \times \Gamma_k)^*$ satisfying the following: there are two positions $i, j \in \{1, \ldots, n\}$ such that $i \sim_u j$, $\bar{b}_i[x] = 1$, and $\bar{b}_j[y] = 1$. Note that L is indeed a regular language so that we can construct a corresponding session automaton $\mathcal{A}$ with k registers such that $L_{symb}(\mathcal{A}) = L$.

In all of the above base cases, if φ is the atomic formula and $\mathcal{A}$ the corresponding session automaton, the following holds: given a data word $\hat{w} \in \mathsf{DW}_k$ encoding free variables, we have $\hat{w} \in L(\varphi)$ iff $\hat{w} \in L(\mathcal{A})$.

Let us come to the induction step. To deal with negation, we can indeed exploit the fact that session automata are closed under complementation when considering only k-bounded data words (Theorem 3.8). Suppose we have constructed a session automaton $\mathcal{A}$ with k registers such that $L(\mathcal{A}) = L(\varphi) \cap \mathsf{DW}_k$. According to Theorem 3.8, there is a session automaton $\mathcal{A}'$ with k registers such that $L(\mathcal{A}') = \mathsf{DW}_k \setminus L(\mathcal{A})$. From $L(\neg\varphi) = (\Sigma \times D)^* \setminus L(\varphi)$, we deduce $L(\mathcal{A}') = L(\neg\varphi) \cap \mathsf{DW}_k$. To deal with disjunction, we exploit closure of session automata under union (again, Theorem 3.8). Finally, existential quantification corresponds, as usual, to projection.

Because of the negations that require complementation (and hence determinization of finite-state automata), the automaton associated with a given dMSO formula has a size given as a tower of exponentials of the size given by the number of nested negations in the formula. □

By Theorems 3.9 and 4.1, we obtain, as a corollary, that model checking session automata against dMSO specifications is decidable, though with non-elementary complexity (while it is undecidable for register automata). Note that this was already shown in [8] for a more powerful model with pushdown stacks.

Theorem 4.2. *Given a session automaton $\mathcal{A}$ and a dMSO sentence φ, one can decide whether $L(\mathcal{A}) \subseteq L(\varphi)$.*

4.2. **Session MSO Logic.** Next, we give an alternative logical characterization of session automata. We identify a fragment of dMSO, called *session MSO logic*, such that every formula from that fragment can be translated into a session automaton, without having to restrict the set of data words in advance. Note that register automata also enjoy a logical characterization [12]. There, *guards* are employed to tame the power of the predicate $\sim$. Similarly, our logic also uses a guard, though in a quite different way.

Definition 4.3. A *session MSO* (sMSO) formula is a dMSO sentence of the form

$$\exists X_1 \cdots \exists X_m \, (\alpha \wedge \forall x \forall y \, (x \sim y \leftrightarrow \beta))$$

such that α and β are classical MSO formulae (not containing the predicate $\sim$).

Example 4.4. The formula $\varphi_1 = \forall x \forall y \, (x \sim y \leftrightarrow x = y)$ is an sMSO formula. Its semantics $L(\varphi_1)$ is the set of data words in which every data value occurs at most once. Moreover, $\varphi_2 = \forall x \forall y \, (x \sim y \leftrightarrow \mathit{true})$ is an sMSO formula, and $L(\varphi_2)$ is the set of data words where all data values coincide. As a last example, let $\varphi_3 = \exists X \, \forall x \forall y \, (x \sim y \leftrightarrow (\neg \exists z \in X \, (x < z \leq y \vee y < z \leq x)))$. Then, $L(\varphi_3)$ is the set of 1-bounded data words. Intuitively, the second-order variable X represents the set of positions where a fresh data value is introduced.

Theorem 4.5. *For all data languages L, the following statements are equivalent:*

- *There is a session automaton $\mathcal{A}$ such that $L(\mathcal{A}) = L$.*
- *There is an sMSO sentence φ such that $L(\varphi) = L$.*

Proof. The construction of an sMSO formula from a session automaton $\mathcal{A}$ has already been sketched in the proof of Theorem 4.1.

We turn to the converse direction and let $\varphi = \exists X_1 \cdots \exists X_m \, (\alpha \wedge \forall x \forall y \, (x \sim y \leftrightarrow \beta))$ be an sMSO sentence. By Theorem 4.1, it is sufficient to show that $L(\varphi)$ is bounded.

The free variables of β are among $x, y, X_1, \ldots, X_m$. As, moreover, β is a "classical" MSO formula, which does not contain $\sim$, it defines, in the expected manner, a set $L_{symb}(\beta)$ of words over the finite alphabet $\Sigma \times \{0,1\}^{m+2}$. Similarly to the proof of Theorem 4.1, the idea is to interpret a position carrying letter $(a, 1, b, b_1, \ldots, b_m)$ as x, and a position labeled $(a, b, 1, b_1, \ldots, b_m)$ as y, while membership in X_i is indicated by b_i. Words where x and y are not uniquely determined, are discarded. We can represent such models as tuples $(w, i, j, I_1, \ldots, I_m)$ where $w \in \Sigma^*$, i denotes the position of the 1-entry in the unique first component, and j denotes the position of the 1-entry in the second component. As $L_{symb}(\beta) \subseteq (\Sigma \times \{0,1\}^{m+2})^*$ is MSO definable (in the classical sense, without data), it is, by Büchi's theorem, recognized by some minimal deterministic finite automaton $\mathcal{A}_\beta$. Suppose that $\mathcal{A}_\beta$ has $k_\beta \geq 1$ states.

We claim that the data language $L(\varphi)$ is k_β-bounded. To show this, let $w = (a_1, d_1) \cdots (a_n, d_n) \in L(\varphi)$. There exists a tuple $\overline{I} = (I_1, \ldots, I_m)$ of subsets of $\{1, \ldots, n\}$ such that, for all $i, j \in \{1, \ldots, n\}$,

$$d_i = d_j \iff (a_1 \cdots a_n, i, j, \overline{I}) \in L_{symb}(\beta) \, . \qquad (*)$$

Suppose, towards a contradiction, that w is not k_β-bounded. Then, there are $k > k_\beta$ and a position $i \in \{1, \ldots, n\}$ such that i is contained in k distinct sessions $J_1, \ldots, J_k$. For $l \in \{1, \ldots, k\}$, let $i_l = \min(J_l)$ and $j_l = \max(J_l)$, so that $J_l = \{i_l, i_l + 1, \ldots, j_l\}$. Note that the i_l are pairwise distinct, and so are the j_l. By $(*)$, for every $l \in \{1, \ldots, k\}$, we have $w_l = (a_1 \cdots a_n, i_l, j_l, \overline{I}) \in L_{symb}(\beta)$. Thus, for every such word w_l, there is a unique accepting run of $\mathcal{A}_\beta$, say, being in state q_l after executing position i. As $\mathcal{A}_\beta$ has only k_β states, there are $l \neq l'$ such that $q_l = q_{l'}$. Thus, there is an accepting run of $\mathcal{A}_\beta$ either on a word where one of the first-order components is not unique, which is a contradiction, or on $(a_1 \cdots a_n, i_l, j_{l'}, \overline{I})$. The latter contradicts $(*)$, since J_l and $J_{l'}$ are distinct sessions. $\square$

5. Learning Session Automata

In this section, we introduce an active learning algorithm for session automata. In the usual active learning setting (as introduced by Angluin [2], see [13] for a general overview of active learning techniques), a *learner* interacts with a so-called minimally adequate *teacher* (MAT), an oracle which can answer *membership* and *equivalence queries*. In our case, the learner is given the task to infer the data language $L(\mathcal{A})$ defined by a given session automaton $\mathcal{A}$. We suppose here that the teacher knows the session automaton or any other device accepting $L(\mathcal{A})$. In practice, this might not be the case — $\mathcal{A}$ could be a black box — and equivalence queries could be (approximately) answered, for example, by extensive testing. The learner can ask if a *data* word is accepted by $\mathcal{A}$ or not. Furthermore it can ask equivalence queries which consist in giving an *hypothesis* session automaton to the teacher who either answers yes, if the hypothesis is equivalent to $\mathcal{A}$ (i.e., both data languages are the same), or gives a data word which is a counterexample, i.e., a data word that is either accepted by the hypothesis automaton but should not, or vice versa.

Given the data language $L(\mathcal{A})$ accepted by a session automaton $\mathcal{A}$ over Σ and D, our algorithm will learn the canonical session automaton $\mathcal{A}^C$, that uses k registers, i.e., the minimal symbolically deterministic automaton recognizing the language $L(\mathcal{A})$ and the regular language $L_{symb}(\mathcal{A}^C)$ over $\Sigma \times \Gamma_k$. Therefore one can consider that the learning target is $L_{symb}(\mathcal{A}^C)$ and use an arbitrary active learning algorithm for regular languages. However, as the teacher answers only questions over data words, queries have to be adapted. Since $\mathcal{A}^C$ only accepts symbolic words which are in normal form, a membership query for a given symbolic word u not in normal form will be answered negatively (without consulting the teacher); otherwise, the teacher will be given one data word included in $\gamma(u)$ (all the answers on words of $\gamma(u)$ are the same). Likewise, before submitting an equivalence query to the teacher, the learning algorithm checks if the current hypothesis automaton accepts symbolic words not in normal form[2]. If yes, one of those is taken as a counterexample, else an equivalence query is submitted to the teacher. Since the number of registers needed to accept a data language is a priori not known, the learning algorithm starts by trying to learn a session automaton with 1 register and increases the number of registers as necessary.

Every active learning algorithm for regular languages may be adapted to our setting. Here we describe a variant of Rivest and Schapire's algorithm [30] which is itself a variant of Angluin's L* algorithm [2]. An overview of learning algorithms for deterministic finite state automata can be found, for example, in [4].

The algorithm is based on the notion of *observation table* which contains the information accumulated by the learner during the learning process. An observation table over a given alphabet $\Sigma \times \Gamma_k$ is a triple $\mathcal{O} = (T, U, V)$ with U, V two sets of words over $\Sigma \times \Gamma_k$ such that $\varepsilon \in U \cap V$ and T is a mapping $(U \cup U \cdot (\Sigma \times \Gamma_k)) \times V \to \{+, -\}$. A table is partitioned into an upper part U and a lower part $U \cdot (\Sigma \times \Gamma_k)$. We define for each $u \in U \cup U \cdot (\Sigma \times \Gamma_k)$ a mapping $row(u) \colon V \to \{+, -\}$ where $row(u)(v) = T(u, v)$. An observation table must satisfy the following property: for all $u, u' \in U$ such that $u \neq u'$ we have $row(u) \neq row(u')$, i.e., there exists $v \in V$ such that $T(u, v) \neq T(u', v)$. This means that the rows of the upper part of the table are pairwise distinct. A table is *closed* if for all $u' \in U \cdot (\Sigma \times \Gamma_k)$ there exists $u \in U$ such that $row(u) = row(u')$. From a closed table we can construct a symbolically

[2]This can be checked in polynomial time over the trimmed hypothesis automaton with a fixed point computation labelling the states with the registers that should be used again before overwriting them.

Algorithm 1: The learning algorithm for a session automaton $\mathcal{A}$

initialize $k := 1$ and $\mathcal{O} := (T, U, V)$ by $U = V = \{\varepsilon\}$ and $T(u, \varepsilon)$ for all $u \in U \cup U \cdot (\Sigma \times \Gamma_k)$ with membership queries;
repeat
  while $\mathcal{O}$ *is not closed* **do**
    find $u \in U$ and $(a, \pi) \in \Sigma \times \Gamma_k$ such that for all $u' \in U$: $row(u(a, \pi)) \neq row(u')$;
    extend table to $\mathcal{O} := (T', U \cup \{u(a, \pi)\}, V)$ by membership queries;
  end
  from $\mathcal{O}$ construct the hypothesized automaton $\mathcal{A}_\mathcal{O}$; `// cf. Definition 5.1`
  if $\mathcal{A}_\mathcal{O}$ *accepts symbolic words not in normal form* **then**
    let z be one of those;
  else
    if $L(\mathcal{A}) = L(\mathcal{A}_\mathcal{O})$ **then**
      equivalence test succeeds;
    else
      get counterexample $w \in (L(\mathcal{A}) \setminus L(\mathcal{A}_\mathcal{O})) \cup (L(\mathcal{A}_\mathcal{O}) \setminus L(\mathcal{A}))$;
      set $z := snf(w)$;
      find minimal k' such that $z \in (\Sigma \times \Gamma_{k'})^*$;
      if $k' > k$ **then**
        set $k := k'$;
        extend table to $\mathcal{O} := (T', U, V)$ over $\Sigma \times \Gamma_k$ by membership queries;
      end
    end
  end
  if $\mathcal{O}$ *is closed* **then** `// is true if` $k' \leq k$
    find a break-point for z where v is the distinguishing word;
    extend table to $\mathcal{O} := (T', U, V \cup \{v\})$ by membership queries;
  end
until *equivalence test succeeds*;
return $\mathcal{A}_\mathcal{O}$

deterministic session automaton whose states correspond to the rows of the upper part of the table:

Definition 5.1. For a closed table $\mathcal{O} = (T, U, V)$ over a finite alphabet $\Sigma \times \Gamma_k$, we define a symbolically deterministic session automaton $A_\mathcal{O} = (S, \mathcal{R}, \iota, F, \Delta)$ over $\Sigma \times \Gamma_k$ by $S = U$, $\mathcal{R} = \{1, \ldots, k\}$, $\iota = \varepsilon$, $F = \{u \in S \mid T(u, \varepsilon) = +\}$, and for all $u \in S$ and $(a, \pi) \in \Sigma \times \Gamma_k$, $\Delta(u, (a, \pi)) = u'$ if $row(u(a, \pi)) = row(u')$. This is well defined as the table is closed.

We now describe in detail our active learning algorithm for a given session automaton $\mathcal{A}$ given in Table 1. It is based on a loop which repeatedly constructs a closed table using membership queries, builds the corresponding automaton and then asks an equivalence query. This is repeated until $\mathcal{A}$ is learned. An important part of a active learning algorithm is the treatment of counterexamples provided by the teacher as an answer to an equivalence query. Suppose that for a given $\mathcal{A}_\mathcal{O}$ constructed from a closed table $\mathcal{O} = (T, U, V)$ the teacher answers by a counterexample data word w. Let $z = snf(w)$. If z uses more registers than available in the current alphabet, we extend the alphabet and then the table. If the obtained table is not closed, we restart from the beginning of the loop. Otherwise – and also if z does not use more registers – we use Rivest and Schapire's [30] technique to extend the table by adding a suitable v to V making it non-closed. The technique is based on the notion of break-point that we now recall. As z is a counterexample, (1) $z \in L_{symb}(\mathcal{A}_\mathcal{O}) \iff z \notin L_{symb}(\mathcal{A}^C)$.

Let $z = z_1 \cdots z_m$, with $z_i \in \Sigma \times \Gamma$. Then, for all i with $1 \leq i \leq m+1$, let z be decomposed as $z = u_i v_i$ with $u_i, v_i \in (\Sigma \times \Gamma)^*$, where $u_1 = v_{m+1} = \varepsilon$, $v_1 = u_{m+1} = z$ and the length of u_i is equal to $i-1$ (we have also $z = u_i z_i v_{i+1}$ for all i such that $1 \leq i \leq m$). Let $s_i \in U$ be the state visited by z just before reading the ith letter, along the computation of z on $\mathcal{A}_{\mathcal{O}}$: i is a break-point if $s_i z_i v_{i+1} \in L_{symb}(\mathcal{A}_{\mathcal{O}}) \iff s_{i+1} v_{i+1} \notin L_{symb}(\mathcal{A}^C)$. Because of (1) such a break-point must exist and can be obtained with $O(\log(m))$ membership queries by a binary search. The word v_{i+1} is called the distinguishing word. If V is extended by v_{i+1} the table is not closed anymore ($row(s_i)$ and $row(s_i z_i)$ become different). Now, the algorithm closes the table again, then asks another equivalence query and so forth until termination. At each iteration of the loop the number of rows (each of those correspond to a state in the automaton $\mathcal{A}^C$) is increased by at least one. Notice that the same counterexample might be given several times. The treatment of the counterexample only guarantees that the table will contain one more row in its upper part. We obtain the following:

Theorem 5.2. *Let $\mathcal{A}$ be a session automaton over Σ and D, using k' registers. Let $\mathcal{A}^C$ be the corresponding canonical session automaton. Let N be its number of states, k its number of registers and M the length of the longest counterexample returned by an equivalence query. Then, the learning algorithm for $\mathcal{A}$ terminates with at most $O(k|\Sigma|N^2 + N\log(M))$ membership and $O(N)$ equivalence queries.*

Proof. This follows directly from the proof of correctness and complexity of Rivest and Schapire's algorithm [4, 30]. Notice that the equivalence query cannot return a counterexample whose normal form uses more than k registers, as such a word is rejected by both $\mathcal{A}^C$ (by definition) and by $\mathcal{A}_{\mathcal{O}}$ (by construction). □

Let us discuss the complexity of our algorithm. In terms of the canonical session automaton, the number of required membership and equivalence queries is polynomial. When the session automaton $\mathcal{A}$ is data deterministic, using the discussion after the proof of Theorem 3.6 over the size of $\mathcal{A}^C$, the overall complexity of the learning algorithm is polynomial in the number of states of $\mathcal{A}$, but exponential in the number of registers it uses (with constant base). As usual, we have to add one exponent when we consider session automata which are not data deterministic. In [19], the number of equivalence queries is polynomial in the size of the underlying automaton. In contrast, the number of membership queries contains a factor n^k where n is the number of states and k the number of registers. This may be seen as a drawback, as n is typically large. Note that [19] restrict to deterministic automata, since classical register automata are not determinizable.

Example 5.3. We apply our learning algorithm on the data language given by the automaton $\mathcal{A}$ of Figure 6(a). In Figure 7 the successive observation tables constructed by the algorithm are given. To save space some letters whose rows contain only $-$'s are omitted. In Figure 8 the successive automata constructed from the closed observation tables are given. For sake of clarity we omit the sink states. We start with the alphabet $\Sigma \times \Gamma_1 = \{(a, 1^{\circledast}), (a, 1^{\uparrow}), (b, 1^{\circledast}), (b, 1^{\uparrow})\}$. We omit letters $(a, 1^{\uparrow})$ and $(b, 1^{\circledast})$. Table $\mathcal{O}_1$ is obtained after initialization and closing by adding $(b, 1^{\uparrow})$ to the top. We use $\_$ to indicate that all letters will lead to the same row. From $\mathcal{O}_1$ the first hypothesis automaton $\mathcal{A}_1$ is constructed. We suppose that the equivalence query gives back as counterexample the data word $(a,3)(b,3)$ whose normal form is $(a, 1^{\circledast})(b, 1^{\uparrow})$. Here the break-point yields the distinguishing word $(b, 1^{\uparrow})$. We add it to V. The obtained table is not closed anymore. We close it by adding $(a, 1^{\circledast})$ to the top and get table $\mathcal{O}_2$ yielding hypothesis automaton $\mathcal{A}_2$.

$\mathcal{O}_1$	ε
ε	+
$(b,1^\uparrow)$	−
$(a,1^\circledast)$	+
$(b,1^\circledast)_-$	−

⇒

$\mathcal{O}_2$	ε	$(b,1^\uparrow)$
ε	+	−
$(b,1^\uparrow)$	−	−
$(a,1^\circledast)$	+	+
$(b,1^\uparrow)_-$	−	−
$(a,1^\circledast)(a,1^\circledast)$	+	+
$(a,1^\circledast)(b,1^\uparrow)$	+	+

⇒

$\mathcal{O}_3$	ε	$(b,1^\uparrow)$
ε	+	−
$(b,1^\uparrow)$	−	−
$(a,1^\circledast)$	+	+
$(a,2^\circledast)$	−	−
$(b,2^\uparrow)$	−	−
$(b,1^\uparrow)_-$	−	−
$(a,1^\circledast)(a,1^\circledast)$	+	+
$(a,1^\circledast)(b,1^\uparrow)$	+	+
$(a,1^\circledast)(a,2^\circledast)$	−	+
$(a,1^\circledast)(b,2^\uparrow)$	−	−

⇒

$\mathcal{O}_4$	ε	$(b,1^\uparrow)$
ε	+	−
$(b,1^\uparrow)$	−	−
$(a,1^\circledast)$	+	+
$(a,1^\circledast)(a,2^\circledast)$	−	+
$(a,2^\circledast)$	−	−
$(b,2^\uparrow)$	−	−
$(b,1^\uparrow)_-$	−	−
$(a,1^\circledast)(a,1^\circledast)$	+	+
$(a,1^\circledast)(b,1^\uparrow)$	+	+
$(a,1^\circledast)(b,2^\uparrow)$	−	−
$(a,1^\circledast)(a,2^\circledast)(a,1^\circledast)$	−	−
$(a,1^\circledast)(a,2^\circledast)(b,1^\uparrow)$	+	+
$(a,1^\circledast)(a,2^\circledast)(a,2^\circledast)$	−	+
$(a,1^\circledast)(a,2^\circledast)(b,2^\uparrow)$	−	+

⇒

$\mathcal{O}_5$	ε	$(b,1^\uparrow)$	$(b,2^\uparrow)$
ε	+	−	−
$(b,1^\uparrow)$	−	−	−
$(a,1^\circledast)$	+	+	−
$(a,1^\circledast)(a,2^\circledast)$	−	+	−
$(a,1^\circledast)(a,2^\circledast)(b,1^\uparrow)$	+	+	+
$(a,2^\circledast)$	−	−	−
$(b,2^\uparrow)$	−	−	−
$(b,1^\uparrow)_-$	−	−	−
$(a,1^\circledast)(a,1^\circledast)$	+	+	−
$(a,1^\circledast)(b,1^\uparrow)$	+	+	−
$(a,1^\circledast)(b,2^\uparrow)$	−	−	−
$(a,1^\circledast)(a,2^\circledast)(a,1^\circledast)$	−	−	−
$(a,1^\circledast)(a,2^\circledast)(a,2^\circledast)$	−	+	−
$(a,1^\circledast)(a,2^\circledast)(b,2^\uparrow)$	−	+	−
$(a,1^\circledast)(a,2^\circledast)(b,1^\uparrow)(a,1^\circledast)$	+	+	+
$(a,1^\circledast)(a,2^\circledast)(b,1^\uparrow)(b,1^\uparrow)$	+	+	+
$(a,1^\circledast)(a,2^\circledast)(b,1^\uparrow)(a,2^\circledast)$	−	+	−
$(a,1^\circledast)(a,2^\circledast)(b,1^\uparrow)(b,2^\uparrow)$	+	+	+

Figure 7: The successive observation tables

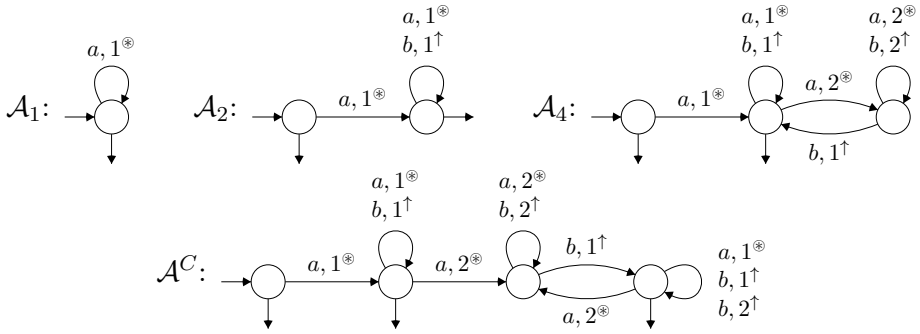


Figure 8: The successive hypothesis automata

Notice that $L_{symb}(\mathcal{A}_2) = L_{symb}(\mathcal{A}^C) \cap (\Sigma \times \Gamma_1)^*$. This means that the equivalence query must give back a data word whose normal form is using at least 2 registers (here $(a,7)(a,4)(b,7)$ with normal form $(a,1^\circledast)(a,2^\circledast)(b,1^\uparrow)$). As the word uses 2 registers, we extend the alphabet to $\Sigma \times \Gamma_2$ and obtain table $\mathcal{O}_3$. We close the table and get $\mathcal{O}_4$. From there we obtain the

hypothesis automaton $\mathcal{A}_4$. After the equivalence query we get $(a, 1^{\circledast})(a, 2^{\circledast})(b, 1^{\uparrow})(b, 2^{\uparrow})$ as normal form of the data word counterexample $(a, 9)(a, 3)(b, 9)(b, 3)$. After adding $(b, 2^{\uparrow})$ to V and closing the table by moving $(a, 1^{\circledast})(a, 2^{\circledast})(b, 1^{\uparrow})$ to the top we get finally the table $\mathcal{O}_5$ from which the canonical automaton $\mathcal{A}^C$ is obtained and the equivalence query succeeds.

6. Conclusion

In this paper, we developed a theory of session automata, which form a robust class of data languages. In particular, they are closed under union, intersection, and resource-sensitive complementation. Moreover, they enjoy logical characterizations in terms of (a fragment of) MSO logic with a predicate to compare data values for equality. Finally, unlike most other automata models for data words, session automata have a decidable inclusion problem. This makes them attractive for verification and learning. In fact, we provided a complete framework for algorithmic learning of session automata, making use of their canonical normal form. An interesting direction to follow would be to try to apply those methods to other models of automata dealing with data values like data automata [6, 5] or variable automata [16]. As a next step, we plan to employ our setting for various verification tasks. In particular, the next step is to implement our framework, using possibly other learning algorithms than the one of Rivest and Shapire that we presented in this article, for instance using the LearnLib platform [27] or libalf [10].

Acknowledgments. We are grateful to Thomas Schwentick for suggesting the symbolic normal form of data words, and to the reviewers for their valuable comments.

References

[1] F. Aarts, F. Heidarian, H. Kuppens, P. Olsen, and F. W. Vaandrager. Automata learning through counterexample guided abstraction refinement. In *FM*, volume 7436 of *Lecture Notes in Computer Science*, pages 10–27. Springer, 2012.

[2] D. Angluin. Learning regular sets from queries and counterexamples. *Information and Computation*, 75(2):87–106, 1987.

[3] T. Berg, O. Grinchtein, B. Jonsson, M. Leucker, H. Raffelt, and B. Steffen. On the correspondence between conformance testing and regular inference. In *FASE*, volume 3442 of *Lecture Notes in Computer Science*, pages 175–189. Springer, 2005.

[4] T. Berg and H. Raffelt. Model checking. In *Model-based Testing of Reactive Systems*, volume 3472 of *Lecture Notes in Computer Science*. Springer, 2005.

[5] H. Björklund and Th. Schwentick. On notions of regularity for data languages. *Theoretical Computer Science*, 411(4-5):702–715, 2010.

[6] M. Bojanczyk, C. David, A. Muscholl, T. Schwentick, and L. Segoufin. Two-variable logic on data words. *ACM Trans. Comput. Log.*, 12(4):27, 2011.

[7] M. Bojańczyk and S. Lasota. An extension of data automata that captures XPath. In *LICS 2010*, pages 243–252. IEEE Computer Society, 2010.

[8] B. Bollig, A. Cyriac, P. Gastin, and K. Narayan Kumar. Model checking languages of data words. In L. Birkedal, editor, *Proceedings of FoSSaCS'12*, volume 7213 of *Lecture Notes in Computer Science*, pages 391–405. Springer, 2012.

[9] B. Bollig, P. Habermehl, M. Leucker, and B. Monmege. A fresh approach to learning register automata. In *Proceedings of the 17th International Conference on Developments in Language Theory (DLT'13)*, volume 7907 of *Lecture Notes in Computer Science*, pages 118–130. Springer, 2013.

[10] B. Bollig, J.-P. Katoen, C. Kern, M. Leucker, D. Neider, and D. Piegdon. libalf: the automata learning framework. In *CAV*, volume 6174 of *Lecture Notes in Computer Science*, pages 360–364. Springer, 2010.

[11] J. M. Cobleigh, D. Giannakopoulou, and C. S. Pasareanu. Learning assumptions for compositional verification. In *TACAS*, volume 2619 of *Lecture Notes in Computer Science*, pages 331–346. Springer, 2003.
[12] T. Colcombet, C. Ley, and G. Puppis. On the use of guards for logics with data. In *Proceedings of MFCS'11*, volume 6907 of *Lecture Notes in Computer Science*, pages 243–255. Springer Berlin / Heidelberg, 2011.
[13] C. de la Higuera. *Grammatical Inference. Learning Automata and Grammars*. Cambridge University Press, 2010.
[14] S. Demri and R. Lazić. LTL with the freeze quantifier and register automata. *ACM Transactions on Computational Logic*, 10(3), 2009.
[15] D. Giannakopoulou and J. Magee. Fluent model checking for event-based systems. In *ESEC / SIGSOFT FSE*, pages 257–266. ACM, 2003.
[16] O. Grumberg, O. Kupferman, and S. Sheinvald. Variable automata over infinite alphabets. In *LATA*, volume 6031 of *Lecture Notes in Computer Science*, pages 561–572. Springer, 2010.
[17] O. Grumberg, O. Kupferman, and S. Sheinvald. An automata-theoretic approach to reasoning about parameterized systems and specifications. In *ATVA*, volume 8172 of *Lecture Notes in Computer Science*, pages 397–411. Springer, 2013.
[18] P. Habermehl and T. Vojnar. Regular model checking using inference of regular languages. *Electronic Notes in Theoretical Computer Science*, 138(3):21–36, 2005.
[19] F. Howar, B. Steffen, B. Jonsson, and S. Cassel. Inferring canonical register automata. In *VMCAI*, volume 7148 of *Lecture Notes in Computer Science*, pages 251–266. Springer, 2012.
[20] B. Jonsson. Learning of automata models extended with data. In *SFM*, volume 6659 of *Lecture Notes in Computer Science*, pages 327–349. Springer, 2011.
[21] M. Kaminski and N. Francez. Finite-memory automata. *Theoretical Computer Science*, 134(2):329–363, 1994.
[22] M. Kaminski and T. Tan. Regular expressions for languages over infinite alphabets. *Fundamenta Informaticae*, 69(3):301–318, 2006.
[23] M. Kaminski and D. Zeitlin. Finite-memory automata with non-deterministic reassignment. *International Journal of Foundations of Computer Science*, 21(5):741–760, 2010.
[24] K. O. Kürtz, R. Küsters, and T. Wilke. Selecting theories and nonce generation for recursive protocols. In P. Ning, V. Atluri, V. D. Gligor, and H. Mantel, editors, *FMSE*, pages 61–70. ACM, 2007.
[25] A. Kurz, T. Suzuki, and E. Tuosto. On nominal regular languages with binders. In L. Birkedal, editor, *Proceedings of FoSSaCS'12*, volume 7213 of *Lecture Notes in Computer Science*, pages 255–269. Springer, 2012.
[26] M. Leucker. Learning meets verification. In *FMCO*, volume 4709 of *Lecture Notes in Computer Science*, pages 127–151. Springer, 2007.
[27] T. Margaria, H. Raffelt, B. Steffen, and M. Leucker. The LearnLib in FMICS-jETI. In *ICECCS*, pages 340–352. IEEE Computer Society Press, 2007.
[28] R. Milner, J. Parrow, and D. Walker. A calculus of mobile processes, Parts I and II. *Information and Computation*, 100:1–77, Sept. 1992.
[29] F. Neven, Th. Schwentick, and V. Vianu. Finite state machines for strings over infinite alphabets. *ACM Transactions on Computational Logic*, 5(3):403–435, 2004.
[30] R. Rivest and R. Schapire. Inference of finite automata using homing sequences. *Information and Computation*, 103:299–347, 1993.
[31] H. Sakamoto and D. Ikeda. Intractability of decision problems for finite-memory automata. *Theoretical Computer Science*, 231:297–308, 2000.
[32] L. Segoufin. Automata and logics for words and trees over an infinite alphabet. In Z. Ésik, editor, *CSL 2006*, volume 4207 of *LNCS*, pages 41–57. Springer, 2006.
[33] N. Tzevelekos. Fresh-register automata. In T. Ball and M. Sagiv, editors, *POPL*, pages 295–306. ACM, 2011.